GLU 2017 International Workshop on Grounding Language Understanding
25 August 2017, Stockholm, Sweden

Language is Not About Language: Towards Formalizing the Role of Extra-Linguistic Factors in Human and Machine Language Acquisition and Communication

Okko Räsänen[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
okko.rasanen@aalto.fi

Abstract

Despite the large research efforts in understanding early language acquisition (LA), it is still unclear how young children learn to transform their noisy and ambiguous auditory experience into a symbolic compositional representational system known as language. This paper argues that a major obstacle towards a more comprehensive picture of LA is the lack of a unified conceptual framework that would capture the full extent of factors critical to language learning in real world contexts, and that we should pursue such a framework in order to be able to place individual behavioral studies and computational models into a mutually compatible context. As an example of the issue, the widely used standard model of the speech chain—a description of the information flow from talker's idea to listener's interpretation of the meaning of the spoken message—is shown to be insufficient for characterizing learning and communication in natural contexts. Instead, a realistic model should account for the inherent multimodality and contextual dependency of communication and learning by formally acknowledging the role of a shared communicative context, interlocutors' subjective representations of the shared situation, and how these factors drive message generation and speech perception in order to acquire information on the external world. By understanding how language is connected to the more generic sensorimotor and predictive processing principles of human cognition, we can also start to understand the core forces driving language learning in natural environments and through varying individual developmental trajectories.

Index Terms: language acquisition; speech perception; contextual grounding; embodiment; speech chain; computational modeling

1. Introduction

Normally developing children learn their native language in a manner that seems almost effortless. Instead of receiving formal instruction, they learn to understand and produce speech simply by interacting with their linguistic environment. Moreover, they do it without any notion of linguistic concepts such as phonemes or syntax that are in the center of academic analysis of languages. However, despite decades of research in several scientific disciplines, *we are still lacking a clear picture of how the early language acquisition (LA) takes place*. For instance, it is still unclear how the discrete and symbolic structure of language emerges from the continuous and noisy perceptual input that learners have access to. In addition, we don't know how the emerging language representations such as phonemes and words are coupled to each other or how they depend on the learning environments, on the quality and quantity of speech input, or on other concurrently developing cognitive factors at different stages of the developmental timeline.

Research on human LA has been paralleled by decades of speech technology research aiming at developing automatic speech recognition (ASR) systems for machines. Despite the numerous advances in the field, including the recent deep learning methods and end-to-end systems trained on large-scale data sets, the state-of-the-art ASR and dialogue systems still fall far behind human performance in natural communication (e.g., [1,2]). This is largely because *ASR systems do not understand speech in a context*, but simply act as speech-to-text transcribers using machine-learning based statistical regression, potentially followed by separate language understanding modules operating on the text output. These systems are also dependent on the availability of labeled training data for the given language or expertise in the use domain—resources that are very expensive to produce. As an alternative solution, some researchers have started to develop so-called zero-resource speech processing systems that could largely operate in an unsupervised manner [3], and hence enable easier deployment of systems in low-resource settings. However, these systems are also currently lacking guiding principles of how languages can be learned without explicit supervision, as exemplified by the low performance of the existing systems [4].

If computational systems and robots could learn similarly to human children, i.e., simply by interacting with their linguistic environments with multiple senses and output channels, this could provide unprecedented scalability, performance, and enable speech applications for the thousands of world's languages and domains where labeled training data is not available, including also non-canonical speakers of high-resourced languages such as people with physical speech impairments (see, e.g., [5]).

However, a major hindrance in understanding both human and machine LA is the lack of a unified conceptual framework that would explain how the bits and pieces of empirical findings, modeling results, and hypothesized underlying cognitive mechanisms contribute to the big picture of language learning and communication. Due to the enormous complexity of language as a phenomenon, scientific reductionism, although necessary, also struggles to integrate individual findings from specialized disciplines and studies back into a coherent whole. In addition, classical views of language having a certain compositional and discrete structure, although useful for characterizing and analyzing languages in general, may be difficult to reconcile with the emerging language skills in children where communication performance, but not analytic competence, has functional

10.21437/GLU.2017-8

significance to a young child. Fragmentation of research and assumptions of linguistic units as real cognitive entities (see also [6]) also leads to the risk of formulating scientific problems or concepts that may not be relevant in the big picture of infant learning, or problems that are at least much easier to solve given the sensorimotor experiences and external and internal constraints available to human learners solving the same problem (see [7] for a case in early word segmentation).

Given this background, the purpose of the present paper is to argue that, in order to truly understand language acquisition, we should work towards formal high-level computational models of language learning (c.f., [8]) that take into account the critical factors behind natural communication and learning beyond the traditional domain of linguistic factors. We should also use the emerging findings from modeling and robotics to feed the theoretical advancements in language research, and update the basic framework of the communicative behavior to better account for the complexity of the reality and the way language is intimately tied to factors beyond its traditional structural descriptions. This also means that we have to go beyond mapping of continuous "noisy" acoustics into abstract linguistic "invariants", but to study how language integrates to the more general cognitive capabilities and perception-action loops the communicating agents. Importantly, I argue that we should place the informative and communicative nature of speech, that is*, capability of language to predict and depict states of the external world*, in the focus instead of directly pursuing linguistic entities as proximal targets of LA.

By formally acknowledging the role of the extra-linguistic factors and characterizing the information flow within the full connected system of agents and their environments, we can start to investigate how different structural descriptions of spoken language (phonemes, words, grammar etc.) serve the communicative and referential purpose of language at different stages of the development, and explore what are the driving forces (c.f., learning criteria in machine learning) behind the emergence of the linguistic capabilities as we witness them in children.

As a simple example, I will sketch out a straightforward extension of a widely used basic model of speech communication, the speech chain, to account for a number of additional dependencies that are likely to play a central role in speech communication and language acquisition. This sketch is not to be taken as complete or its separation into sub-components as properly determined. Instead, the purpose is simply to demonstrate some of the aspects that are fully absent from the basic models of the production-perception chain and why such a model is not suitable for understanding language learning.

2. Extending the speech chain

I believe that a large part of the so-called "unification problem" in language research originates from the combination of scientific reductionism with a commonly adopted view of language as a structured system that is largely independent of the external world and of the embodied agents engaged in communication. This so-called *traditional view,* one that I here characterize in an intentionally stereotypical manner to make a point, is a view that has been largely influenced by the structuralist tradition in linguistics. The view is potentially most clearly reflected in phonology where relation to speech acoustics has long remained underspecified (e.g., [9]). However, similar linguistic concepts, originally meant as *descriptive*, have been adopted to related fields dealing with language research such as psychology and neuroscience.

According to the traditional view, it is still obvious that language could not exist or function without the world beyond the language. For instance, it is acknowledged that the contextual grounding is needed to explain the semantics of language symbols (c.f., [10, 11]), that communicative behavior cannot be understood without considering the shared context and inter-agent relationships (e.g., [12, 13]), and that languages evolve through use in different social and societal contexts, to name a few examples.

However, the core structure of the language, that is, the hierarchical compositional structure consisting of units such as phonemes, morphemes, words, grammar etc. is often considered to be a self-contained abstract system that exists independently of the external world or personal (embodied) experiences of the interlocutors (e.g., [14]). Instead of being mere descriptors of language organization at multiple levels—descriptors that have been developed to suit human researchers' limited capability to analytically deal with continuous spaces and probabilistic phenomena of sensorimotor language—such structural units are often taken as the ground truth for spoken language out there in the wild, and hence also often considered as central targets for speech perception and language learning in humans and machines.

To give a concrete example, the structure-focused and context-independent view of language is most clearly reflected in a widely utilized model of speech communication, namely the concept of a speech chain (see, e.g., [15]; Fig. 1): Starting from an idea in the brain of a talker, this idea is formulated into a linguistic message, articulated into acoustic speech heard by a listener, who then attempts to decode the linguistic message from the noisy signal and finally infers its correct meaning. It is said that the talker informs the listener, and this chain of events can be mathematically characterized in terms of the fidelity of the decoded message (e.g., KL-divergence $D_{KL}\{C \,\|\, C^*\}$ in the notation of Fig. 1).

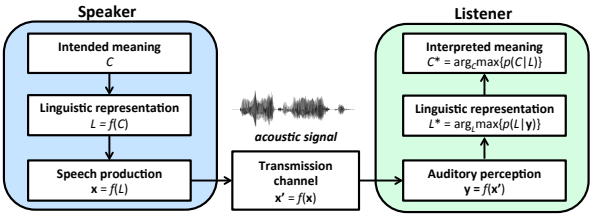

Figure 1: *Standard speech chain model from talker's intentions to listener's interpretation of the meaning.*

The speech chain model is an intuitively appealing description of adult communication and therefore very pervasive in language research and speech engineering. In the study of LA, the linguistic decoding process is typically further divided into multiple sub-tasks that the infant supposedly has to solve (e.g., phonemic/syllable perception, word segmentation, word recognition, syntactic analysis, prosodic analysis, inference of meaning etc.) and that are then studied and modeled individually (but see, e.g., [7, 16–18] for recent joint-models and references to the previous single-process models). Although it is well known that learning of word meanings requires grounding of word forms to their external referents, other processing stages are typically treated as sequential steps between the perceived acoustic input and the ultimately

decoded meaning, closely following the steps of the speech chain. In general, it is often assumed that the language-to-be-learned is a closed symbolic system with a certain internal structure, and infants simply need to infer the correct rules and regularities from the observed acoustic speech, potentially supported by a number of internal biases.

While the standard speech chain is useful as a first approximation, it has turned out to be very difficult to find satisfactory explanations to how infants could infer discrete linguistic representations from realistic speech input where temporal and categorical boundaries between different linguistic constituents are mixed with non-phonological acoustic variation, and when infants receive no formal instruction in their native language but simply have the desire to communicate.

Importantly, the standard speech chain does not consider the following facts: 1) *communication is nearly always dependent on the context (concrete or abstract)* 2) *the human brain has evolved to model and predict the surrounding environment* (e.g., [19, 20], and 3) *language is all about communicating about the surrounding world and the other agents in it.* The standard model assumes that: **A1)** the original idea-to-be-communicated emerges out of nowhere, **A2)** the receiver is already capable of transforming acoustic speech to a linguistic code and then to the intended idea, and **A3)** the perception and comprehension processes are independent of the production process and of the ongoing situation. For a communication scenario between a language learning infant and a caregiver, *none of these three assumptions are true*.

First, all word meanings emerge from the associations between acoustic word forms and their *external referents* whose *internal representations* co-occur with these words. Born without any language-specific knowledge (c.f., **A2**), the only way to start learning meaningful language patterns is to *perceive speech in the context of active internal representations for the external world, obtained through perception*. This also means that, in order for any grounding to take place, the *messages produced by the caregiver cannot be independent of the shared environment* but are statistically coupled to it (c.f., **A1**). Finally, speech comprehension—a process that is always measured in terms of the resulting *behavior* in the same context—is also constrained by the likely meanings available in the situation, and the behavior is also the only component that ultimately matters for the agent (c.f., **A3**). This suggests that children's communication performance can far exceed their formal linguistic competence, functional communication preceding maturation of the language representations available to literate adults.

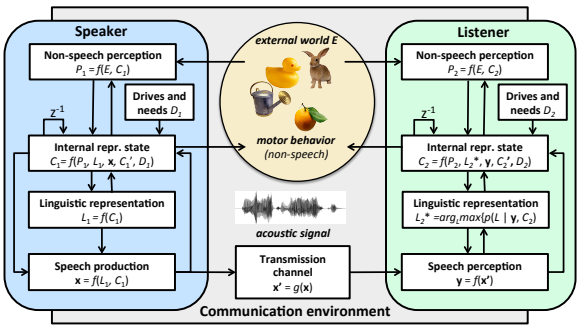

Figure 2: *Extended speech chain that takes into account the shared communicative context of the two interacting agents, providing the basis for contextual grounding of language.*

In general, the internal representations for perceived acoustic speech and for the external (referential) world *must interact* in the processing chain or otherwise speech would not convey any information at all. This very fundamental aspect of language is fully absent from the standard speech chain model, as are any other explanatory factors for, e.g., interaction dynamics (e.g., [21] or top-down effects such as experience-based selective attention in language processing (e.g., [22]).

2.1. Accounting for critical additional factors

If we wish to use a model analogous to the speech chain for capturing the basic components relevant to LA, the model should be extended to account for the situatedness as well as cross-modal and top-down/bottom-up interactions known to play an important role in human cognition. Fig. 2 shows a schematic view of a potential extension to the speech chain where a number of critical components are added to the basic model:

1) **A shared context** (incl. other agents) that can be perceived and manipulated by the interlocutors through a number of sensors and motor channels. This is the domain that connects language patterns to the states and events in the external world that are ecologically relevant to the learner, i.e., *the source of the meanings*.

2) **Internal representational states and transformations of the agents** (e.g., working memory contents & primed associations, current needs and intentions, emotional state etc.) that actually form the content and interactions of linguistic and non-linguistic patterns in the environment and within the agent, including translation of speech input into more general embodied sensorimotor representations of its "meaning". This is the *domain of meanings* where language alters non-linguistic activity of the brain and vice versa.

3) **The effects of present internal representational states on the ongoing perceptual processing and motor activity**. These links are needed to account for the well-documented top-down predictive behavior, selective attentional mechanisms, and general selective sampling and learning from the input [19, 20, 23, 24] that enable efficient learning in ambiguous environments with partial data and finite sensing and processing resources, but where humans also tend to deviate from ideal observers and learners with infinite processing resources without time constraints.

4) **Innate drives and needs** of the agents. Without an internal mechanism rewarding for exploration and learning, no interesting behaviors such as LA would take place (see, e.g., [25] for a review). Such mechanisms have to be understood if we wish to understand, e.g., the dynamics of infant-caregiver interaction and how it generates language data suitable for the developmental stage of the child. Young children do not try to learn language for its own sake, but learning results from exposure to rewarding situations.

By including such factors, phenomena such as interaction, joint-attention, and context-guided speech production and perception become concrete properties of the system dynamics instead of being "something extra" on top of message passing. Importantly, by making the interactions and dependencies visible, it becomes possible to develop mathematical models of information flow in such a system, characterize how different latent (structural) representations of input can impact this information flow, and ultimately identify what are the critical factors driving the learning process and ensuring convergence to a functioning communication system despite

the varying language experiences of individual learners. Due to the computational formalism, these models can be always translated into functional computational implementations [8] and tested with real world or carefully simulated data.

For instance, a hypothesis arising from a fully connected model of grounded communication is that an efficient representation for spoken language becomes defined as the one that provides maximal information regarding the external world given the observed acoustic patterns and learnability constraints of the system. This is very different from the traditional structuralist idea of first learning linguistic patterns and rules from sensory input – an underdetermined problem (see, e.g., [7]) – and only later associating these patterns with representations that play a functional role for the agent.

3. Discussion and related work

The inclusion of a referential context into models of LA is hardly a new idea, as the research during the last 20 years has shown that the constraints from communicative contexts and interactions between agents can explain emergence of various functional communicative representations that are hard to acquire based on speech acoustics alone.

For instance, the early robotic CELL model by Roy and Pentland [26] and the following model by Yu and Ballard [27] both utilize cross-modal regularities in order to find phoneme sequences that consistently co-occur with certain visual referents. Computational models developed in ACORNS EU FP6 FET project used (simulated) visual context to bootstrap word learning from continuous speech in the absence of any a priori notion of phonemic or lexical entities (e.g., [28–32]; see also [33] for an overview). The robotic system described by Salvi et al. [34] also learns words and their meanings by detecting recurring co-occurrences between utterances and the concurrent actions and perceived environments. Recently, Räsänen and Rasilo [7] also provided information-theoretic motivation and empirically validated computational results for an idea that segmentation of speech into words – a task that has been extensively studied as a separate stage of language acquisition – is in fact unnecessary for bootstrapping language learning as long as word learning is treated as a contextually grounded process already from the start.

In terms of speech production, there have also been a variety of models showing how the system of native speech sounds can be learned in interaction with linguistically proficient caregivers without imposing strong a priori constraints on the number and type of categories that exist in the language (e.g., [35, 36]; see also [37] for a review). A number of robotic studies also demonstrate how new functional communicative systems and representations can emerge from situated interaction, internal drive for exploration, and capability to observe and act upon the environment beyond "linguistic" messaging (e.g., [38–40]). Finally, the referential nature of communication is also becoming utilized in models attempting to explain the structuring of sound systems [41].

In general, it is theoretically necessary that some type of contextual grounding is needed during language acquisition, minimally to connect words to their referents in the external world. However, the constantly growing literature in human and machine LA (see [7] for a summary on the former), only briefly scratched above, suggests that the communicative context may play a bigger role than acting a source of meanings from pre-segmented and phonemically represented auditory word forms. In addition, it can be argued that there is no direct ecological pressure for an infant to acquire an abstract symbolic system such as language, unless this system is also somehow related to the ability to comprehend and predict the world, i.e., to guide the perception-action loop, avoid harm, and provide rewards. Hence, we should acknowledge the role of external world in our conceptions of speech communication, especially during early language acquisition where speech is initially void of any meaning.

As for the other briefly listed factors such as the differences between external world and the agent-specific internal representations of it, active attention-driven sensing of the environment, and the interaction of bottom-up percepts and top-down predictions from the generic cognitive machinery, large amounts of work have already been conducted on all these topics separately and also in the context of language, and are beyond any meaningful review here. However, understanding how all the different factors work in concert to produce meaningful communication skills is not yet known, and this is where formal computational models could be used to bridge the gaps between empirical details and sub-process models with the learning behavior at large and across different developmental stages.

4. Conclusions

Research in robotics and computational modeling are at the frontier in pursuing better understanding of language as a phenomenon. This is because functional systems operating with real world sensory data provide an excellent platform for both quantifying the computational principles underlying successful acquisition of communication skills, language learning, and for testing the plausibility of theoretical models in practice. Importantly, computational models are also especially suited to integrate individual research findings and process theories into unified systems, as all computational models have to explicitly specify the inputs, outputs, and all information processing steps in the system.

The argument of the present paper is that, in order to pursue more comprehensive understanding of early language learning, we should bravely seek to combine the universality of computational descriptions with the modern understanding of language as an embodied and contextually grounded process that taps into generic information processing principles of the mammalian brain. By including the critical "non-linguistic" factors to the basic conceptions and frameworks used to guide our thinking and work in language research, we can also start to ask correct questions, evaluate our models in a more meaningful manner, and, importantly, integrate findings on different aspects of language processing into unified coherent descriptions that are compatible with empirical data across disciplinary boundaries. Given the modern-day opportunities to create and evaluate increasingly comprehensive models of LA with access to rich real-world data, we should explore the extent that extralinguistic factors and parallel developing cognitive functions play a role in the emergence of language-related representations in the minds of the learners, and thereby go beyond proximal inference of phonemes and lexemes as targets of language acquisition.

5. Acknowledgements

This research was funded by the Academy of Finland project titled "Computational Modeling of Language Acquisition" (no. 274479).

6. References

[1] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 34–43, 2012.

[2] B. Meyer, H. Hermansky, and N. Morgan, "Bio-inspired speech recognition: From human perception to deep networks," A tutorial held at *Interspeech-2015*, September 6, 2015, Dresden, Germany.

[3] J. Glass, "Towards unsupervised speech processing," *Proc. Int. Conf. Information Science, Signal Processing and their Applications (ISSPA),* Montreal, Canada, 2012, pp. 1–4.

[4] M. Versteegh, R. Thiollière, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015," *Proc. Interspeech-2015*, Dresden, Germany, 2015, pp. 3169–3173.

[5] B. Ons, J. Gemmeke, and H. Van hamme, "Fast vocabulary acquisition in an NMF-based self-learning vocal interface," *Computer Speech & Language*, vol. 28, pp. 997–1017, 2014.

[6] R. Port and A. Leary, "Against formal phonology," *Language*, 81, 927–964, 2005.

[7] O. Räsänen and H. Rasilo, "A joint model of word segmentation and meaning acquisition through cross-situational learning," *Psychological Review*, vol. 122, pp. 792–829, 2015.

[8] D. Marr, *Vision: A Computational Approach*. San Francisco, Freeman & Co, 1982.

[9] J. Ohala, "There is no interface between phonology and phonetics: a personal view," *Journal of Phonetics,* vol. 18, pp. 153–171, 1990.

[10] W. Quine, *Word and Object*. MIT Press, 1960.

[11] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, pp. 335–346, 1990.

[12] H. P. Grice, "Logic and conversation," In P. Cole & J. Morgan (Eds.): *Syntax and Semantics. 3: Speech Acts*. New York: Academic Press (pp. 41–58), 1975.

[13] H. Clark and T. Carlson, "Context for comprehension," In J. Long & A. Baddeley (Eds.): *Attention and performance IX.* Hillsdale, N.J.: Erlbaum, 1981.

[14] J. Katz and P. Postal. *An integrated theory of linguistic description.* Cambridge, Massachusetts: MIT Press, 1964.

[15] P. Denes and E. Pinson, *The speech chain*: *the physics and biology of spoken language*. New York, N.Y.: W.H. Freeman, 1993.

[16] A. Fourtassi and E. Dupoux, "A rudimentary lexicon and semantics help bootstrap phoneme acquisition," *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*, Baltimore, Maryland, 2014, pp. 191–200.

[17] M. Elsner, S. Goldwater, N. Feldman, and F. Wood, "A joint model of word segmentation, lexical acquisition, and phonetic variability," *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-2013)*, Seattle, Washington, 2013, pp. 42–54.

[18] N. Feldman, T. Griffiths, S. Goldwater, and J. Morgan, "A role for the developing lexicon in phonetic category acquisition," *Psychological Review*, vol. 120, pp. 751–778, 2013.

[19] K. Friston, "The free-energy principle: a unified brain theory?", *Nature Reviews Neuroscience*, vol 11, pp. 127–138, 2010.

[20] K. Friston and S. Kiebel, "Cortical circuits for perceptual inference," *Neural Networks*, vol. 22, pp. 1093–1104, 2009.

[21] P. Kuhl, F.-M. Tsao, and H.-M. Liu, "Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning," *PNAS*, vol. 100, pp. 9096–9101, 2003.

[22] A. Zarcone, M. van Schijndel, J. Vogels, and V. Demberg, "Salience and attention in surprisal-based accounts of language processing," *Frontiers in Psychology*, vol. 7, article 844, 2016.

[23] D. Yurovsky, L. Smith, and C. Yu, "Statistical word learning at scale: The baby's view is better," *Developmental Science*, vol. 16, pp. 959–966, 2013.

[24] K. Federmeier, "Thinking ahead: The role and roots of prediction in language comprehension," *Psychophysiology*, vol. 44, pp. 491–505, 2007.

[25] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends in Cognitive Sciences*, vol. 17, pp. 585–593, 2013.

[26] D. Roy and A. Pentland, A. P. "Learning words from sights and sounds: A computational model," *Cognitive Science,* vol. 26, pp. 113–146, 2002.

[27] C. Yu and D. Ballard, "A multimodal learning interface for grounding spoken language in sensory perceptions," *ACM Transactions on Applied Perception*, vol. 1, pp. 57–80, 2004.

[28] O. Räsänen, U. Laine, and T. Altosaar, "Computational language acquisition by statistical bottom-up processing," *Proc. Interspeech-2008,* Brisbane, Australia, 2008, pp. 1980–1983.

[29] G. Aimetti, "Modelling early language acquisition skills: Towards a general statistical learning mechanism", *Proc. Student Research Workshop at the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, 2009, pp. 1–9.

[30] H. Van hamme, "HAC-models: A novel approach to continuous speech recognition," *Proc. Interspeech-2008*, Brisbane, Australia, 2008, pp. 2554–2557.

[31] L. ten Bosch, H. Van hamme, L. Boves, L., and R. Moore, "A computational model of language acquisition: The emergence of words," *Fundamenta Informaticae,* vol. 90, pp. 229–249, 2009.

[32] O. Räsänen and U. Laine, "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences," *Pattern Recognition*, vol. 45, pp. 606–616, 2012.

[33] O. Räsänen, "Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions," *Speech Communication*, vol. 54, pp. 975–997, 2012.

[34] G. Salvi, L. Montesano, A. Bernardino, and J. Santos-Victor, "Language bootstrapping: Learning word meanings from perception-action association," *IEEE Transactions on Systems, Man, and Cybernetics Part B, Cybernetics,* vol. 42, pp. 660–671, 2012.

[35] I. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, pp. 85–117, 2011.

[36] H. Rasilo and O. Räsänen, "An online model of vowel imitation learning," *Speech Communication*, vol. 86, pp. 1–23, 2017.

[37] M. Asada and N. Endo, "Infant-caregiver interactions affect the early development of vocalizations," *Proc. Annual Int. Conf. of the IEEE Engineering in Medicine and Biological Society*, Milan, Italy, 2015, pp. 5351–5354.

[38] L. Steels, "Language Games for Autonomous Robots", *IEEE Intelligent Systems*, September/October 2001, pp. 16–22, 2001.

[39] L. Steels and F. Kaplan, "AIBO's first words: The social learning of language and meaning," *Evolution of Communication*, vol. 4, pp. 3–32, 2000.

[40] P.-Y. Oudeyer, "The self-organization of speech sounds," *Journal of Theoretical Biology*, vol. 233, pp. 435–449, 2005.

[41] C. Moulin-Frier, J. Diard, J-L. Schwartz, and P. Bessière, "COSMO ("Communicating about objects using sensory-motor operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems," *Journal of Phonetics*, vol. 53, pp. 5–41, 2015.